

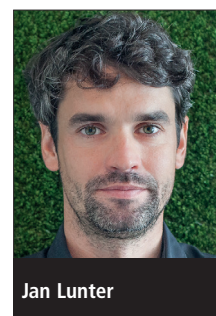


Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Beating the bias in facial recognition technology

Jan Lunter, Innovatrics



In 2019, San Francisco became the first US city to ban facial recognition technology (FRT), specifically vetoing its use by police and other agencies¹. Since then, several other American cities have implemented their own similar FRT bans, with Boston's city councillors² explicitly highlighting one particular issue: the technology's bias.

In addition, moves to prohibit FRT have come not only from public officials, but the corporate world as well. In a widely circulated open letter in June³, IBM CEO Arvind Krishna outlined several proposals to promote racial justice, including the fact that "IBM no longer offers general-purpose IBM facial recognition or analysis software". Emphasising his point, Krishna added: "Vendors and users of AI systems have a shared responsibility to ensure that AI is tested for bias, and that such bias testing is audited and reported."

IBM's share price hit a three-month high that day, nearly matching pre-pandemic levels. The public accolade for taking internal action against FRT and suggesting long-term alternative solutions was a boon to the company – even though its facial recognition department was in fact unaffected: commentators noted⁴ that this was a "symbolic decision" unlikely to affect IBM's bottom line, as the company had already removed facial detection from its API in September 2019⁵.

Trust a two-way street

For IBM, dropping facial recognition to focus on other sectors was a workaround to the general problem of bias in AI systems. Yet while this company may have temporarily stepped back from this field, FRT remains a cornerstone of our AI-powered future – with applications ranging from systems to reduce the spread of Covid-19 by limiting physical contact, to software that expedites the identification process in airports, public buildings, places of employment, and anywhere else where trust and safety are paramount. There's no abandoning a future in which FRT works for us all.

Clearly, the successful and widespread adoption of FRT depends not only on its speed and accuracy, but also on the public's trust of the algorithms that power facial recognition devices. And that trust will be difficult to build while examples of racial bias and false positive results continue to overshadow the year-by-year

improvements being made in the software's overall accuracy and functionality. This challenge must be addressed, to ensure the adoption of FRT remains driven by its ability to provide better security for users worldwide.

"Public trust in facial recognition will be difficult to build while examples of racial bias and false positive results continue to overshadow the year-by-year improvements being made in its accuracy and functionality"

Identifying the problem

IBM's performance as a facial recognition provider, prior to its abandonment of the market in summer 2020, is difficult to assess. One 2017 study found that IBM's facial recognition algorithm had an 87% success rate, compared to Microsoft's success rate of 93.7%, and 90% for Face++, the first online facial recognition platform in China⁶.

Yet these results, while backed by data and rigorously tested, still don't give us the best

assessment of facial recognition performance. IBM, together with Amazon – the other key provider of FRT to the US police – did not submit their algorithms to the National Institute of Standards and Technology (NIST), which sets the standard for commercial use of AI in America. NIST's ongoing Face Recognition Vendor Tests (FRVTs) have evaluated over 400 facial recognition algorithms since 2017, giving insight into the overall accuracy ratings of all the participating developers⁷.

NIST recently ran a large-scale test focused on identifying bias in FRT, with a particular emphasis on the false positive rate – ie, the frequency with which an algorithm misidentifies one person's image. The results showed that "across demographics, false positive rates often vary by factors of 10 to beyond 100 times", depending on which algorithms are in use⁸. The most accurate algorithms produced significantly fewer errors, highlighting just how crucial quality is when choosing FRT.

This data also shone a light on the presence of racial bias. NIST found that even the best algorithms still displayed a higher false positive rate among West and East African and East Asian individuals, while Eastern Europeans had the lowest false positive rate. In short, the tested algorithms tended to mis-identify photos of Asian and Black individuals more than they misidentified Caucasians. The researchers also pointed out⁹: "It is commonly accepted that



one of the major sources of performance differential in modern face recognition engines based on deep convolutional neural networks is demographic imbalances in the training data used to train these engines.”

The NIST study, published in December 2019, made clear the existence of the bias problem in FRT. But to solve it, vendors and users alike need to know where the solutions lie – and recognise that the problem may not be as straightforward as it seems. As a recent study of commercial facial recognition algorithms led by Mei Wang¹⁰ showed: “All algorithms and APIs perform the best on Caucasian testing subsets, followed by Indian, and the worst on Asian and African. This is because the learned representations predominantly trained on Caucasians will discard useful information for discerning non-Caucasian faces.” The researchers also pointed out the importance of representation of faces in the dataset: “APIs which are developed by East Asian companies (Baidu) perform better on Asians, while APIs developed in the Western hemisphere perform better on Caucasians.”

Right ID approach

In fact, not all algorithms analysed in the 2019 NIST study performed relatively worse at identifying East Asians. The study found that FRT systems developed in China tend to have a low false positive rate when it comes to authenticating East Asians. This highlights the crucial importance of datasets in algorithm testing. NIST, for example, uses photos sourced from visa photos, mugshots, pictures taken at the US border and similar images for its algorithms to identify¹¹. And Chinese companies participating in these tests simply have more photos of East Asian individuals than companies working outside that continent.

So the case for diversifying the datasets used in algorithm development is clear, but it doesn't tell the full story. The cross-race effect comes into play here – that is, the tendency for individuals to more correctly discern faces of their own race. This propensity has been well-documented, particularly in the context of law enforcement, where individuals may be asked to identify someone of a different race in a line-up¹².

A 2001 analysis of police cases found that cross-racial identifications carried out by humans were correct a mere 46% of the time – far below even the least-accurate facial recognition algorithms¹³. In short, it's important to note that while reducing bias in FRT remains a key priority, the technology already significantly outperforms witness-based methods of identification.

There are other notable challenges on the path toward a facial recognition algorithm that is 100% accurate. For example, darker skin tones

reflect less light, and therefore provide less detail for facial recognition algorithms to analyse. As the above-mentioned study by Mei Wang *et al* pointed out: “Even with balanced training, we see that non-Caucasians still perform more poorly than Caucasians. The reason may be that faces of coloured skin are more difficult to extract and pre-process feature information, especially in dark situations.” This challenge, while hardly insurmountable, has slowed the development of accurate facial recognition for people with darker skin tones. But luckily, there are long-term solutions to these issues already in place.

Accuracy for all

Wide-scale and standardised studies like NIST's help to uncover common flaws in FRT. As both scientists and employees, the engineers who develop FRT are tasked with improving the software's results in recognising images with darker skin tones – as well as reducing the gender gap in false positive rates, which was also noted in the 2019 NIST study.

When the datasets and results are put to rigorous, minute analysis, scientists have the ability to improve the algorithm and thereby address the issue of bias in future iterations of the software. This direct approach puts an emphasis on the issues in order to solve them, although it can also over-exaggerate the problem when reported to the general public. When higher rates of false positives are found for certain demographics – be they race, gender or otherwise – facial recognition companies can approach the problem using the insights that the biometrics industry has gained over the past two decades. These include:

- Better data labelling. Modern facial recognition algorithms are the products of machine learning and neural networks, which analyse millions of annotated images to learn how to discern faces. If these datasets are poorly labelled, certain groups of people will be more difficult to recognise. What's more, a self-training neural network will accept mislabelled data as fact, thereby embedding the error within the system. To improve rather than worsen the situation, FRT algorithms require rich, varied datasets that are double and triple-checked as a standardised priority.

- External dataset auditing. Unbiased datasets make for unbiased algorithms. With the increased presence of FRT in our daily lives, it is more crucial than ever to ensure datasets are properly and independently audited, in order to reduce bias. The alternative is to show that the dataset has been deemed as balanced. For example, Yaobin Zhang and Weihong Deng recently built a class-balanced dataset from public image data used for facial recognition training¹⁴. As they pointed out: “Our pub-

licly available dataset is characterised by the uniformly distributed sample size per class, as well as the balance between the number of classes and the number of samples in one class. Experimental results show that deep models trained with the BUPT-CBFace dataset can not only achieve comparable results to larger-scale datasets such as MS-Celeb-1M, but also alleviate the problem of recognition bias.”

- Reducing algorithmic bias. While this approach is relatively new, it is one of the solutions that shows most promise. For example, Alexander Amini and his colleagues showed that by using de-biasing variational auto-encoders, they were able to automatically discover and mitigate hidden biases among the training data. As they said in their research paper¹⁵: “We tackle the challenge of integrating de-biasing capabilities directly into a model training process that adapts automatically and without supervision to the shortcomings of the training data. Our approach features an end-to-end deep learning algorithm that simultaneously learns the desired task (eg, facial detection) as well as the underlying latent structure of the training data. Learning this latent distribution in an unsupervised manner enables us to uncover hidden or implicit biases within the training data.”

With an increased emphasis on non-biased algorithms, the use of such de-biasing may become commonplace. This should also not incur a performance hit, as the researchers demonstrated that their de-biasing approach provided “increased overall performance as well as decreased categorical bias”.

- Removing undetected duplicates. While false positives are more common, algorithms can also create false negatives – the failure to discern the same person in two different pictures. This can be due to an appearance change or some difference in the photo quality. In either case, by reviewing datasets for duplicate, low-quality and other unsuitable data, developers can improve accuracy and reduce bias.

Looking ahead

Between 2014 and 2018, the accuracy of facial recognition technology increased 20-fold¹⁶. And the continued improvements we will see over the next several years will likely bring myriad new uses for FRT – but new challenges too. The rapid and constant improvement of algorithms will enable bias to be reduced far below the current levels. Driven by stringent testing standards, and the need to gain an edge over the competition, errors that could be perceived as biases are being rapidly assessed, managed and refined. Better algorithms reduce bias by improving accuracy, leading to better results for the vendor company.



Police using FRT could fine-tune the system so it is less accurate overall, but the chance of bias is reduced: if an algorithm is unsure, it will simply register an error which can be checked, rather than risk creating a false positive.

This is an iterative process. And while the pace of technological development in biometrics has been notable, companies seeking to reduce bias faster than the current rate of facial recognition progress do have some options here. One method is to fine-tune the algorithm in such a way that although the algorithm is less accurate overall, it reduces the chances of bias. Simply put, if an algorithm is unsure, it will just register an error rather than risk the chance – however small – of creating a false positive.

Such an approach could be appropriate for companies working in sectors that are subject to high public scrutiny, such as law enforcement or government. Reducing the overall accuracy of the algorithm is a cost-effective way to allow the FRT to process images in bulk, with any errors being subject to a manual check for extra safety. In this way, agencies still have a resource to use that reduces the chance of mis-identification, while continuing to build public trust in an algorithm that does what it can with what it's been given.

However, this approach may not remain widely acceptable for long. There is clearly an accelerated need for fast, efficient and now contactless identification methods. But in the current global recession, any reduction in overall FRT effectiveness will be met with scrutiny by buyers who need the best solution for their budgets.

Bans like those enacted in San Francisco and Boston can buy time for officials, buyers and public opinion to decide where they stand on the issue of bias in FRT. For the companies supplying this technology, the choice is either to get out of the game entirely, as IBM did – or ensure their algorithms are tested to the highest standards on diverse, accurately labelled datasets that have been either de-biased or certified that they are not skewed towards any particular gender or race.

Luckily, the current shortcomings of the technology can and are being investigated by researchers, resulting in algorithms that are able to spot hidden biases. Any failure to use these techniques will not only fan public mistrust, but also inhibit the iterative pace of improvement

shown over the past five years. It will be vital for FRT developers to communicate the improvements we will see over the next five years in order to fulfil both the potential and purpose of facial recognition, no matter who the user is.

About the author

Jan Lunter is the co-founder and CEO of Innovatrics, which has been developing and providing fingerprint recognition solutions since 2004. He is also author of a fingerprint analysis and recognition algorithm that regularly ranks among the top in prestigious comparison tests (NIST PFT II, NIST Minex). In recent years, Jan has also focused on image processing and the use of neural networks for face recognition. He graduated from the Télécom ParisTech University in France.

References

1. Kate Conger, Richard Fausset and Serge F Kovalski. 'San Francisco Bans Facial Recognition Technology', New York Times, 14 May 2019. Accessed September 2020. <https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html?auth=login-email&login=email>.
2. Ally Jarmaning. 'Boston Bans Use Of Facial Recognition Technology'. WBUR News, 24 June 2020. Accessed September 2020. <https://www.wbur.org/news/2020/06/23/boston-facial-recognition-ban>.
3. 'IBM CEO's Letter to Congress on Racial Justice Reform'. IBM, 8 June 2020. Accessed September 2020. <https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/>.
4. Matt O'Brien. 'IBM quits facial recognition, joins call for police reforms'. AP News, 9 June 9 2020. Accessed September 2020. <https://apnews.com/5ee4450df46d2d96bf85d7db683bb0a6>.
5. 'Visual Recognition: Release notes'. IBM. Accessed September 2020. <https://cloud.ibm.com/docs/visual-recognition?topic=visual-recognition-release-notes>.
6. 'Gender Shades AI project'. Algorithmic Justice League. Accessed September 2020. <http://gendershades.org/overview.html>.
7. 'FRVT 1:1 Verification Project'. NIST. Accessed September 2020. https://pages.nist.gov/frvt/html/frvt11.html#_frvt_participation_statistics_.
8. Patrick Grother, Mei Ngan and Kayee Hanaoka. 'Face Recognition Vendor Test Part 3: Demographic Effects'. NIST, December 2019. Accessed September 2020. <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf>.
9. Martins Bruveris, Pouria Mortazavian, Jochem Gietema and Mohan Mahadevan. 'Reducing Geographic Performance Differentials for Face Recognition', Onfido UK. Accessed September 2020. https://openaccess.thecvf.com/content_WACVW_2020/papers/w1/Bruveris_Reducing_Geographic_Performance_Differentials_for_Face_Recognition_WACVW_2020_paper.pdf.
10. Mei Wang, Weihong Deng and Jiani Hu, Beijing University of Posts and Telecommunications; and Xunqiang Tao and Yaohai Huang, Canon Information Technology (Beijing) Co Ltd. 'Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network'. 27 July 2019. Accessed September 2020. <https://arxiv.org/pdf/1812.00194.pdf>.
11. 'FRVT 1:1 Verification Report'. NIST. Accessed September 2020. https://pages.nist.gov/frvt/html/frvt11.html#_frvt_participation_statistics_.
12. Kathleen L Hourihan, Memorial University of Newfoundland; Aaron S Benjamin, University of Illinois at Urbana-Champaign; and Xiping Liu, Tianjin Normal University. 'A cross-race effect in metamemory: Predictions of face recognition are more accurate for members of our own race'. 2 July 2012. Accessed September 2020. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3496291/>.
13. B W Behrman and S L Davey. 'Eyewitness identification in actual criminal cases: An archival analysis'. APA PsycNet Journal, 2001. Accessed September 2020. <https://doi.apa.org/doiLanding?doi=10.1023%2FA:1012840831846>.
14. Yaobin Zhang and Weihong Deng, Beijing University of Posts and Telecommunications. 'Class-Balanced Training for Deep Face Recognition'. Accessed September 2020. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w48/Zhang_Class-Balanced_Training_for_Deep_Face_Recognition_CVPRW_2020_paper.pdf.
15. Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia and Daniela Rus, Massachusetts Institute of Technology. 'Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure'. January 2019. Accessed September 2020. https://www.researchgate.net/publication/334381622_Uncovering_and_Mitigating_Algorithmic_Bias_through_Learned_Latent_Structure.
16. 'NIST Evaluation Shows Advance in Face Recognition Software's Capabilities'. NIST, 30 November 2018. Accessed September 2020. <https://www.nist.gov/news-events/news/2018/11/nist-evaluation-shows-advance-face-recognition-software-capabilities>.